On the Convergence of Monte Carlo Maximum Likelihood Calculations

By
Charles J. Geyer[1].
Technical Report No. 571
School of Statistics
University of Minnesota
February 18, 1992

[1]Research supported in part by grant DMS-9007833 from the National Science Foundation

Abstract

Monte Carlo maximum likelihood (Geyer and Thompson, 1992) can be used for an extremely broad class of models. Given any family $\{ h_\theta : \theta \in \Theta \}$ of nonnegative integrable functions, maximum likelihood estimates in the family obtained by normalizing the the functions to integrate to one can be done using Monte Carlo maximum likelihood, the only regularity conditions being that the evaluation maps $\theta \mapsto h_\theta(x)$ be lower semicontinuous for almost all x and upper semicontinuous for the observed x. The precise result is that under these conditions the Monte Carlo approximant to the log likelihood hypoconverges to the exact log likelihood. This implies convergence of maximizers, of profile likelihoods, and of level sets of the likelihood. The same result is obtainable when there are missing data (Gelfand and Carlin, 1991), but a Wald-type integrability condition needs to be imposed, the integrability being with respect to the conditional distribution of the missing data given the observed data. Conditions for asymptotic normality are also discussed.

1 Normalized Families of Densities

Suppose we have a family of nonnegative functions

$$\{ h_\theta : \theta \in \Theta \}$$

on a probability space, all of which are integrable with respect to a measure μ and none integrating to zero. Let the integrals be denoted

$$c(\theta) = \int h_\theta \, d\mu$$

Then for each θ in Θ the function f_θ defined by

$$f_\theta(x) = \frac{1}{c(\theta)} h_\theta(x)$$

is a probability density with respect to μ. We we call a family $\{ f_\theta : \theta \in \Theta \}$ of this form a *normalized family* of densities. The function $\theta \mapsto c(\theta)$ is the *normalizer* of the family, and the functions h_θ are the *unnormalized densities* or the *predensities* of the family. We denote the distribution corresponding to θ by P_θ,

$$P_\theta(A) = \int_A f_\theta(x) \, d\mu(x)$$

for any measurable set A, and expectation with respect to P_θ by E_θ,

$$E_\theta g(X) = \int g(x) f_\theta(x) \, d\mu(x)$$

for any integrable function g.

Such families are interesting because for arbitrary functions h_θ realizations X_1, X_2, ... can be simulated without knowledge of the normalizer $c(\theta)$ by the Hastings algorithm (Hastings, 1970). Moreover, maximum likelihood estimation can be carried out, again without knowledge of the normalizer or its derivatives, using these Monte Carlo simulations (Geyer and Thompson, 1992).

The log likelihood corresponding to an observation x we take for convenience to be the likelihood ratio against an arbitrary fixed parameter point ψ

$$\begin{aligned} l(\theta) &= \log \frac{h_\theta(x)}{h_\psi(x)} - \log \frac{c(\theta)}{c(\psi)} \\ &= \log \frac{h_\theta(x)}{h_\psi(x)} - \log E_\psi \frac{h_\theta(X)}{h_\psi(X)} \end{aligned} \tag{1}$$

since

$$E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x) \, d\mu(x) = \frac{1}{c(\psi)} \int h_\theta(x) \, d\mu(x) = \frac{c(\theta)}{c(\psi)}. \tag{2}$$

It is not actually necessary that $\psi \in \Theta$, only that P_ψ dominate P_θ for all $\theta \in \Theta$ so that the set of points x such that $h_\psi(x) = 0$ can be ignored in the integrals in (2).

Given a sample $X_1, \ldots, X_n$ from P_ψ generated by the Hastings algorithm, the natural Monte Carlo approximation of the log likelihood is

$$l_n(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)} \tag{3}$$

where $E_{n,\psi}$ denotes the 'empirical' expectation with respect to P_ψ defined by

$$E_{n,\psi} g(X) = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

If the Markov chain $X_1, X_2, \ldots$ generated by the Hastings algorithm is irreducible, then $E_{n,\psi} g(X)$ converges almost surely to $E_\psi g(X)$ for any integrable function g. In particular, for any fixed θ, $l_n(\theta)$ converges almost surely to $l(\theta)$. The 'almost surely' here means for almost all sample paths of the Hastings algorithm. We are treating the observation x as fixed. Only the simulations $X_1, X_2, \ldots$ are treated as random. Note that the nullset of sample paths for which convergence fails may depend on θ.

Let $\hat{\theta}$ be the maximizer (assumed unique for a moment) of the true log likelihood l and let $\hat{\theta}_n$ be an ϵ_n-maximizer of l_n

$$l_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} l_n(\theta) - \epsilon_n$$

for some sequence $\{\epsilon_n\}$ converging to zero.

Geyer and Thompson (1992) show that if the normalized family in question is an exponential family, i. e. $h_\theta(x) = \exp(\langle t(x), \theta \rangle)$, then $\hat{\theta}_n$ converges to $\hat{\theta}$ for almost all sample paths of the Monte Carlo simulation. The proof relies on the fact that log likelihoods of exponential families are concave. Geyer and Thompson remark that an analogous result should hold outside of exponential families. The next section gives such a theorem.

2 Likelihood Convergence

2.1 Set Convergence

At several points we will need the concept of Painlevé-Kuratowski set convergence (Sec 3A in Rockafellar and Wets, forthcoming; Sec 1.4.1 in Attouch, 1984). Given a sequence of sets C_n, the set limit superior is the set of points x such that there is a subsequence $x_{n_k} \to x$ with $x_{n_k} \in C_{n_k}$, and the set limit inferior is the set of points x such that there is a sequence $x_n \to x$ with $x_n \in C_n$ for all n after some n_0. If the set limsup and liminf agree, then their common value is said to be the limit of the sequence.

Another characterization that is valid only in a locally compact topological space (e. g. $\mathbf{R}^d$) uses the so-called 'hit or miss' criteria (Proposition 3A.10 in Rockafellar and Wets, forthcoming; Theorem 2.75 in Attouch, 1984). A set C is the limit of the sequence C_n if and only if C_n eventually hits every open set that hits C and

eventually misses every compact set that misses C, i. e. for every open set O such that $O \cap C$ is nonempty and every compact set K such that $K \cap C$ is empty, there is an m such that $O \cap C_n$ is nonempty and $K \cap C_n$ is empty for all $n \geq m$.

2.2 Epiconvergence and Hypoconvergence

Epiconvergence and hypoconvergence are types of convergence of sequences of functions that are useful in optimization problems. If a sequence of functions g_n epiconverges to a limit g (written $g_n \xrightarrow{e} g$) and x_n minimizes g_n then any cluster point of the sequence $\{x_n\}$ is a minimizer of g. Hypoconvergence is the analogous notion for maximization problems. Since x maximizes g if and only if it minimizes $-g$, hypoconvergence (written $g_n \xrightarrow{h} g$) is defined by $g_n \xrightarrow{h} g$ if and only if $(-g_n) \xrightarrow{e} (-g)$.

A sequence of functions g_n *epiconverges* to a function g if the following two conditions hold (Attouch, 1984, p. 30)

(a) for every point x and for every sequence $x_n \to x$

$$\liminf_n g_n(x_n) \geq g(x).$$

(b) for every point x there exists a sequence $x_n \to x$ such that

$$\limsup_n g_n(x_n) \leq g(x).$$

Epiconvergence is a combination of one-sided uniform convergence, with something weaker than pointwise convergence from the other side. Condition (a) is the one-sided uniform convergence, and condition (b) follows from pointwise convergence (take $x_n \equiv x$) but does not imply it. An equivalent pair of conditions are the following (Attouch, 1984, p. 26). For every point x

$$g(x) \leq \sup_{B \in \mathcal{N}(x)} \liminf_{n \to \infty} \inf_{y \in B} g_n(y) \tag{4a}$$

$$g(x) \geq \sup_{B \in \mathcal{N}(x)} \limsup_{n \to \infty} \inf_{y \in B} g_n(y) \tag{4b}$$

where $\mathcal{N}(x)$ denotes the set of neighborhoods of the point x. The right hand side of (4a) is called the epi-limit inferior and the right hand side of (4b) the epi-limit superior, but we will not need this terminology.

Another characterization of epiconvergence that is sometimes taken as a definition uses the notion of the *epigraph* of a function $g: S \to \mathbf{R}$, the set

$$\operatorname{epi} g = \{ (x, \lambda) \in S \times \mathbf{R} : g(x) \leq \lambda \}$$

of points in $S \times \mathbf{R}$ lying on or above the graph. A sequence of functions g_n epiconverges to a function g if and only if the sequence of sets $\operatorname{epi} g_n$ converges to the set $\operatorname{epi} g$. This gives us 'hit or miss' criteria for epiconvergence. A sequence of functions g_n on a locally compact topological space S epiconverges to a function g if and only

if the sequence epi g_n eventually hits every open set in $S \times \mathbf{R}$ that hits epi g and eventually misses every compact set K in $S \times \mathbf{R}$ that misses epi g.

The main reason for the importance of epiconvergence is the following proposition, which is Theorem 1.10 in Attouch (1984).

Proposition 1 *Suppose* $g_n \xrightarrow{e} g$, $x_n \to x$ *and*

$$g_n(x_n) - \inf g_n \to 0$$

(i. e. x_n *is an* ϵ_n*-minimizing sequence), then*

$$g(x) = \inf g = \lim_{n \to \infty} g_n(x_n)$$

That is, if x_n is an ϵ_n-minimizer of g_n, then any convergent subsequence of $\{x_n\}$ must converge to a point x which minimizes g and the optimal values $g_n(x_n)$ must also converge to the asymptotic optimal value $g(x)$. Two points are worth comment here. First, there is no requirement that the minimizers be unique. If g has a unique minimizer x, then x is the only cluster point of the sequence $\{x_n\}$. Otherwise, there may be many cluster points, but all of them must minimize g. Second, the proposition does not rule out escape to infinity; it only describes what happens if $x_n \to x$. It does say that if the sequence $\{x_n\}$ is confined to a compact set and if g has a unique minimizer, then x_n converges to that minimizer.

2.3 Hypoconvergence of the Monte Carlo Likelihood

Let us now specialize these results to Monte Carlo likelihood. Since we are maximizing rather than minimizing we want to show that the Monte Carlo log likelihood (3) hypoconverges to the true log likelihood (1).

Theorem 1 *For a normalized family of densities determined by unnormalized densities* $\{ h_\theta : \theta \in \Theta \}$ *indexed by a parameter set* Θ *which is a second countable topological space (e. g.,* $\mathbf{R}^2$*), if the evaluation maps* $\theta \mapsto h_\theta(x)$ *are lower semicontinuous for all* x *except a* P_ψ *nullset and upper semicontinuous for the observed* x *and if the Hastings algorithm is irreducible, then the Monte Carlo log likelihood (3) hypoconverges to the true log likelihood (1) with probability one. Also the true log likelihood is upper semicontinuous and the normalizer of the family is lower semicontinuous.*

PROOF. What is to be shown is the hypoconvergence equivalent of (4)

$$l(\theta) \leq \inf_{B \in \mathcal{N}(\theta)} \liminf_{n \to \infty} \sup_{\varphi \in B} l_n(\varphi) \tag{5a}$$

$$l(\theta) \geq \inf_{B \in \mathcal{N}(\theta)} \limsup_{n \to \infty} \sup_{\varphi \in B} l_n(\varphi) \tag{5b}$$

By assumption there is a countable base $\mathcal{B}$ for the topology of Θ. Hence Θ also has a countable dense set Θ_c (just take a point in each member of $\mathcal{B}$). We will need

$$\lim_{n \to \infty} E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)} = E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)} \tag{6}$$

and

$$\lim_{n\to\infty} E_{n,\psi} \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} = E_\psi \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} \tag{7}$$

to hold simultaneously for all $\theta \in \Theta_c$ and all $B \in \mathcal{B}$. This follows from the irreducibility assumption, since the union of a countable number of nullsets (one exception set for each limit) is still a nullset.

First we tackle (5a). If $B \in \mathcal{B}$ and $\theta \in B \cap \Theta_c$

$$l(\theta) = \lim_{n\to\infty} l_n(\theta) \le \liminf_{n\to\infty} \sup_{\varphi\in B} l_n(\varphi)$$

by (6). So

$$\sup_{\varphi\in B\cap\Theta_c} l(\varphi) \le \liminf_{n\to\infty} \sup_{\varphi\in B} l_n(\varphi)$$

and

$$\inf_{B\in\mathcal{N}(\theta)} \sup_{\varphi\in B\cap\Theta_c} l(\varphi) \le \inf_{B\in\mathcal{N}(\theta)} \liminf_{n\to\infty} \sup_{\varphi\in B} l_n(\varphi)$$

The left hand side is equal to $l(\theta)$ if and only if l is upper semicontinuous. Hence upper semicontinuity of l implies (5a). Since $\theta \mapsto h_\theta(x)$ is assumed to be upper semicontinuous for the observed x and since a sum of functions is upper semicontinuous if both functions are, it remains only to be shown that $-\log[c(\theta)/c(\psi)]$ is upper semicontinuous, which is true if the normalizer $c(\theta)$ is lower semicontinuous, which follows from Fatou's lemma and the lower semicontinuity of $\theta \to h_\theta(x)$: if $\theta_k \to \theta$

$$c(\theta) = \int \left(\liminf_{k\to\infty} h_{\theta_k}(x)\right) d\mu(x) \le \liminf_{k\to\infty} \int h_{\theta_k}(x)\, d\mu(x) = \liminf_{k\to\infty} c(\theta_k)$$

This establishes (5a) and the assertions about upper and lower semicontinuity of the log likelihood and the normalizer.

Now

$$\begin{aligned}
\inf_{B\in\mathcal{N}(\theta)} \limsup_{n\to\infty} \sup_{\varphi\in B} l_n(\varphi) &\le \inf_{B\in\mathcal{N}(\theta)} \left(\sup_{\varphi\in B} \frac{h_\phi(x)}{h_\psi(x)} - \liminf_{n\to\infty} \inf_{\varphi\in B} \log E_{n,\psi} \frac{h_\phi(X)}{h_\psi(X)}\right) \\
&= \frac{h_\theta(x)}{h_\psi(x)} - \log \sup_{B\in\mathcal{N}(\theta)} \lim_{n\to\infty} E_{n,\psi} \inf_{\varphi\in B} \frac{h_\phi(X)}{h_\psi(X)} \\
&\to \frac{h_\theta(x)}{h_\psi(x)} - \log \sup_{B\in\mathcal{N}(\theta)} E_\psi \inf_{\varphi\in B} \frac{h_\phi(X)}{h_\psi(X)}
\end{aligned}$$

where the equality follows from the upper semicontinuity of $\theta \mapsto h_\theta(x)$ and the limit follows from (7). The limit will be equal to $l(\theta)$ and establish (5b) if

$$\sup_{B\in\mathcal{N}(\theta)} E_\psi \inf_{\varphi\in B} \frac{h_\phi(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)}$$

Now the integrand here satisfies

$$0 \le \inf_{\varphi\in B} \frac{h_\varphi(x)}{h_\psi(x)} \le \frac{h_\theta(x)}{h_\psi(x)}, \qquad \forall\, x$$

(since $\theta \in B$). Since the right hand side is integrable by (2) and the evaluation maps are assumed lower semicontinuous, dominated convergence implies

$$\sup_{B \in \mathcal{N}(\theta)} E_\psi \inf_{\varphi \in B} \frac{h_\varphi(X)}{h_\psi(X)} \to E_\psi \sup_{B \in \mathcal{N}(\theta)} \inf_{\varphi \in B} \frac{h_\varphi(X)}{h_\psi(X)} = E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)}$$

(The apparent uncountable sup over the whole neighborhood filter $\mathcal{N}(\theta)$ is the same as the sup over the countable neighborhood base $\mathcal{N}(\theta) \cap \mathcal{B}$.) This completes the proof. □

2.4 Convergence of the MLE Calculation

Corollary 1 *If a sequence $\{\hat{\theta}_n\}$ of ϵ_n-maximizers of the Monte Carlo log likelihood is bounded (i. e. contained in a compact set) almost surely (resp. in probability) and there is a unique maximum likelihood estimate $\hat{\theta}$, then $\hat{\theta}_n \to \hat{\theta}$ almost surely (resp. in probability).*

PROOF. The assertion about almost sure convergence follows directly from the theorem and Proposition 1. If $\{\hat{\theta}_n\}$ is contained in a compact set, then every subsequence has a convergent subsubsequence, and each such subsubsequence must converge to $\hat{\theta}$. Hence the whole sequence converges to $\hat{\theta}$.

The assertion about convergence in probability follows by almost the same argument. A sequence bounded in probability is tight, hence every subsequence has a subsubsequence which converges in distribution by Prohorov's theorem. By Skorohod representation, the convergence can be considered almost sure, in which case the only possible limit is $\hat{\theta}$. Hence the whole sequence converges in distribution to the point mass at $\hat{\theta}$, which is the same as convergence in probability to $\hat{\theta}$. □

The corollary applies trivially when the whole parameter space Θ is a compact set. This is the usual way in which proofs of this sort proceed, following Wald (1949), who used the one-point compactification, Kiefer and Wolfowitz, (1956), who used more general compactifications, and Bahadur (1971), who gives a very general formulation, showing that most models are compactifiable in the appropriate topology (the one induced by vague convergence of the associated probability measures). Lacking a suitable compactification, it would be necessary to establish by ad hoc methods a uniform upper bound on the sup of l_n outside some very large ball.

2.5 Convergence of Profile Likelihoods and Level Sets

Suppose Θ is a subset of $\mathbf{R}^d$ and θ can be divided $\theta = (\phi, \eta)$ into a 'parameter of interest' ϕ and a 'nuisance parameter' η ($\phi = (\theta_1, \ldots, \theta_k)$ and $\eta = (\theta_{k+1}, \ldots, \theta_d)$). Then the profile likelihood for ϕ is

$$l_p(\phi) = \sup_\eta l\Big((\phi, \eta)\Big)$$

Corollary 2 *If the Monte Carlo log likelihood hypoconverges to the true log likelihood, then any Monte Carlo profile log likelihood also hypoconverges to the true profile log likelihood.*

PROOF. For this we use the 'hit or miss' criteria. Fix an open set O in $\mathbf{R}^k$ that hits the hypograph of l_p. Then the preimage of O under the projection

$$\{ (\phi, \eta) \in \Theta : \phi \in O, \ \eta \in \mathbf{R}^{d-k} \}$$

is open in $\mathbf{R}^d$ and hits the hypograph of l. Hence the preimage is eventually hit by the hypograph of l_n, which implies that O is eventually hit by the hypograph of $l_{n,p}$. Now fix a compact set set K in $\mathbf{R}^k$ that misses the hypograph of l_p. Then for any value η of the nuisance parameter, the set

$$K_\eta = \{ (\phi, \eta) \in \Theta : \phi \in K \}$$

is compact in $\mathbf{R}^d$ and misses the hypograph of l, which implies that K_η is eventually missed by the hypograph of l_n. Since this holds for all η, the whole preimage of K under the projection is eventually missed by the hypograph of l_n, which implies that K is eventually missed by the hypograph of $l_{n,p}$. □

It is perhaps worth a remark that none of the linear structure of $\mathbf{R}^d$ was used in the proof, only its structure as a locally compact group under addition. It would be enough for Θ to be a subset of a locally compact group G with a subspace H and homogeneous space G/H, the parameter of interest being the projection on G/H. (The need for local compactness arises from the use of the hit or miss criteria).

Hypoconvergence also implies a type of convergence for level sets of the of the log likelihood

$$\operatorname{lev}_\alpha l = \{ \theta : l(\theta) \geq \alpha \}$$

This is a direct consequence of Theorem 3.C11 in Rockafellar and Wets (forthcoming), which we state here as follows.

Proposition 2 *A sequence of functions l_n on a locally compact space hypoconverges to l, if and only if both of the following conditions hold*

(a) for every sequence $\alpha_n \to \alpha$

$$\limsup_n (\operatorname{lev}_{\alpha_n} l_n) \subset \operatorname{lev}_\alpha l$$

(b) for some sequence $\alpha_n \to \alpha$

$$\liminf_n (\operatorname{lev}_{\alpha_n} l_n) \supset \operatorname{lev}_\alpha l$$

The useful conclusion of the theorem is (b), since we want to contain the true level set. The 'for some sequence' in (b) is not useful, since we have no way to determine the sequence. These considerations lead to the following.

Corollary 3 *If the Monte Carlo log likelihood hypoconverges to the true log likelihood, then*

$$\limsup_n \operatorname{lev}_\alpha l_n \subset \operatorname{lev}_\alpha l$$
$$\liminf_n \operatorname{lev}_\alpha l_n \supset \operatorname{lev}_\beta l, \qquad \beta > \alpha,$$

and if

$$\overline{\bigcup_{\beta>\alpha} \operatorname{lev}_\beta l} = \operatorname{lev}_\alpha l, \tag{8}$$

also holds, then

$$\lim_n \operatorname{lev}_\alpha l_n = \operatorname{lev}_\alpha l. \tag{9}$$

PROOF. The first assertion is a direct consequence of the proposition. The second follows from the nesting of level sets and the fact that set liminfs (and limsups) are closed ($\operatorname{lev}_\alpha l$ is closed because a hypo-limit is always upper semicontinuous). □

Instead of looking at a fixed level α, we might instead look at a fixed distance γ down from the maximum, i. e., we might consider the sets

$$S_{n,\gamma} = \{\, \theta \in \Theta : l_n(\theta) \le l_n(\hat{\theta}_n) - \gamma \,\}$$

and

$$S_\gamma = \{\, \theta \in \Theta : l(\theta) \le \sup l - \gamma \,\}.$$

But this brings up the question of whether $l_n(\hat{\theta}_n) \to \sup l$. It does under the conditions of Corollary 1, but need not otherwise. So we state a corollary with this as a condition (the proof is the same).

Corollary 4 *If the Monte Carlo log likelihood hypoconverges to the true log likelihood, and if $l_n(\hat{\theta}_n) \to \sup l$, then*

$$\limsup_n S_{n,\gamma} \subset S_\gamma$$
$$\liminf_n S_{n,\gamma} \supset S_\delta, \qquad \delta > \gamma,$$

and if (8) also holds for $\alpha = \sup l - \gamma$, then

$$\lim_n S_{n,\gamma} = S_\gamma$$

Before leaving the subject of likelihood convergence it is perhaps worth pausing for a moment and comparing the results obtained here with the results that are obtainable for the exponential family case (Geyer, 1990, Geyer and Thompson, 1992). The log likelihood and Monte Carlo log likelihood for an exponential family are concave, and this has two consequences that improve the preceding results. First, the boundedness assumptions of Corollary 1 are unnecessary. If concave functions hypoconverge to a concave function that has no directions of recession, then the sequence is *equi-level-bounded*, that is, eventually dominated by a function with compact level sets (Rockafellar and Wets, forthcoming, Propositions 3C.21 and 3C.22). Hence the sequence of Monte Carlo MLEs is eventually contained in any of these compact level sets for levels below $l(\psi)$. The second difference is that (8) is true for any concave function for any level below the maximum (Rockafellar, 1970, Theorem 7.6). So (9) holds automatically for $\alpha < \sup l$.

3 Missing Data

Gelfand and Carlin (1991) have proposed an extension of the methods described above to the case of missing data. Suppose that the random variable X for which h_θ gives densities is not observed, but just some function of it X_{obs}. Then the log likelihood, obtained by integrating over the missing data is

$$l(\theta) = \log E_\psi \left(\frac{h_\theta(X)}{h_\psi(X)} \,\middle|\, X_{obs} \right) - \log E_\psi \frac{h_\theta(X)}{h_\psi(X)} \tag{10}$$

and its natural Monte Carlo approximation is

$$l_n(\theta) = \log E_{n,\psi} \left(\frac{h_\theta(X)}{h_\psi(X)} \,\middle|\, X_{obs} \right) - \log E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)}, \tag{11}$$

where $E_{n,\psi}$, as before, denotes an average over samples from P_ψ generated by the Hastings algorithm and $E_{n,\psi}(\,\cdot\,|X_{obs})$ denotes an average over a second set of samples generated by another Hastings algorithm simulating the conditional distribution of X given X_{obs}. Gelfand and Carlin suggest maximizing (11) to obtain an approximation to the MLE.

If we attempt to apply the program of the last section, we find it doesn't work without additional assumptions. As before, the second term in (11) hypoconverges to the second term in (10), and the same program applied to the first term shows that the first term in (11) *epiconverges* to the first term in (10), but that doesn't do us any good. We need some control on the supremum of the first term uniformly on compact sets, and a dominated convergence argument won't give such control, since the assumptions of Theorem 1 don't imply a dominating function.

So to get a theorem we need to impose a Wald-type integrability condition following Wald (1949). This gives uniform convergence for the first term and hypoconvergence for the second term, which implies hypoconvergence for the sum (Exercise 3D.8, in Rockafellar and Wets, forthcoming). This gives the following theorem. The proof is omitted, since it is just a combination of the proof of Theorem 1 with the methods of Wald (1949) along the lines just described.

Theorem 2 *For a normalized family of densities determined by unnormalized densities $\{ h_\theta : \theta \in \Theta \}$ indexed by a parameter set Θ which is a second countable topological space, if the evaluation maps $\theta \mapsto h_\theta(x)$ are lower semicontinuous semicontinuous for all x except a P_ψ nullset and upper semicontinuous for all x except a $P_\psi(\,\cdot\,|X_{obs})$ nullset, if the Hastings algorithm is irreducible, and if for every $\theta \in \Theta$ there is a neighborhood B of θ such that*

$$E_\psi \left(\sup_{\varphi \in B} \frac{h_\phi(X)}{h_\psi(X)} \,\middle|\, X_{obs} \right) < \infty$$

then the Monte Carlo log likelihood (10) hypoconverges to the true log likelihood (11) with probability one.

Since Corollaries 1, 2, and 3 used only hypoconvergence, they apply to the missing data case as well.

4 A Central Limit Theorem

In contrast to mere convergence, for which the required conditions are very weak, a central limit theorem for $\sqrt{n}(\hat{\theta}_n - \hat{\theta})$ is problematical. We can easily copy one of the usual proofs for the asymptotics of maximum likelihood, making the appropriate changes. But the resulting regularity conditions are not easy to verify, except in special cases.

Theorem 3 *Suppose the following assumptions hold*

(a) The MLE $\hat{\theta}$ is unique and the parameter space Θ contains an open neighborhood of $\hat{\theta}$ in $\mathbf{R}^d$.

(b) The Monte Carlo MLE $\hat{\theta}_n$ converges in probability to $\hat{\theta}$.

(c) $c(\theta) = \int h_\theta \, d\mu$ can be differentiated twice under the integral sign.

(d) $\sqrt{n} \nabla l_n(\hat{\theta}) \xrightarrow{\mathcal{L}} N(0, A)$ for some covariance matrix A.

(e) $B = -\nabla^2 l(\hat{\theta})$ is positive definite.

(f) $\nabla^3 l_n(\theta)$ is bounded in probability uniformly in a neighborhood of $\hat{\theta}$.

then

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}) \xrightarrow{\mathcal{L}} N(0, B^{-1} A B^{-1})$$

A proof would be entirely classical and is omitted.

All of the conditions except (d) are fairly straightforward, and one can imagine verifying them (if they hold) by standard methods. The matrix B in condition (e) cannot be calculated analytically, but $-\nabla^2 l_n(\hat{\theta})$ is a consistent estimate under these conditions. Condition (e) and be verified using dominated convergence and ergodicity if an integrable function can be found that dominates third partial derivatives with respect to theta of h_θ / h_ϕ.

Condition (d) is hard, if Markov chain Monte Carlo is being used for the simulations, because it involves a Markov chain central limit theorem. General Markov chain central limit theorems do exist (Nummelin, 1984), but they seem difficult to apply in practice. Some work in this direction has been done in the context of Markov chain Monte Carlo (Shervish and Carlin, 1990; Chan, 1991; Liu, Wong and Kong, 1991; Tierney, 1991), but it seems that it is difficult to show that a central limit theorem holds for practical models in which the sample space is not finite. If the sample space is finite, the central limit theorem is classical (see, for example, Chung, 1967, p. 99 ff.)

Even assuming that (d) holds, the variance A cannot be calculated using available theory and must be estimated by Monte Carlo.

$$\begin{aligned} \nabla l_n(\theta) &= \frac{\nabla h_\theta(x)}{h_\theta(x)} - \frac{E_{n,\psi} \frac{\nabla h_\theta(X)}{h_\psi(X)}}{E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)}} \\ &= \frac{E_{n,\psi} \left[\left(t_\theta(x) - t_\theta(X) \right) \frac{h_\theta(X)}{h_\psi(X)} \right]}{E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)}} \end{aligned} \tag{12}$$

where $t_\theta(X) = \nabla h_\theta(X)/h_\psi(X)$. Using assumption (c) to differentiate under the integral sign

$$\begin{aligned}\nabla l(\theta) &= \frac{\nabla h_\theta(x)}{h_\theta(x)} - \frac{\nabla c(\theta)}{c(\theta)} \\ &= \frac{\nabla h_\theta(x)}{h_\theta(x)} - \int \frac{\nabla h_\theta(x)}{h_\theta(x)} \frac{h_\theta(x)}{c(\theta)}\, d\mu(x) \\ &= t_\theta(x) - E_\theta t_\theta(X),\end{aligned}$$

and this is zero when $\theta = \hat{\theta}$. The denominator in (12) converges to $c(\theta)/c(\psi)$; the expectation of the numerator with respect to P_ψ is

$$\begin{aligned}E_\psi\big(t_\theta(x) - t_\theta(X)\big)\frac{h_\theta(X)}{h_\psi(X)} &= \frac{c(\theta)}{c(\psi)} \int \big(t_\theta(x) - t_\theta(y)\big) f_\theta(y)\, d\mu(y) \\ &= \frac{c(\theta)}{c(\psi)}\big(t_\theta(x) - E_\theta t_\theta(X)\big),\end{aligned}$$

which is also zero when $\theta = \hat{\theta}$. Thus the numerator is the sample mean for a functional of the Markov chain

$$z_\theta(X) = \big(t_\theta(x) - t_\theta(X)\big)\frac{h_\theta(X)}{h_\psi(X)}$$

which has expectation zero under the stationary distribution. Hence by the continuous mapping theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_\theta(X_i) \xrightarrow{\mathcal{L}} \frac{c(\psi)}{c(\theta)} N(0, A)$$

Then, if the $z_\theta(X_i)^2$ are uniformly integrable,

$$\frac{1}{n} E\left(\sum_{i=1}^{n} z_\theta(X_i)\right)^2 \to \sum_{t=-\infty}^{+\infty} \gamma(t)$$

where $\gamma(t) = \gamma(-t)$ is the lag t autocovariance of the chain at stationarity, i. e.

$$\gamma(t) = \mathrm{Cov}\left(z_\theta(X_0), z_\theta(X_t)\right)$$

if the starting position X_0 of the Markov chain is a realization from P_ψ. Hence

$$A = \frac{c(\theta)^2}{c(\psi)^2} \sum_{t=-\infty}^{+\infty} \gamma(t)$$

both terms of which can be estimated, $c(\theta)/c(\psi)$ by the denominator in (12), and the sum by standard time series methods as

$$\sum_{t=-(n-1)}^{(n-1)} w(t)\hat{\gamma}(t)$$

where

$$\hat{\gamma}(t) = \hat{\gamma}(-t) = \frac{1}{n}\sum_{i=1}^{n-t} z_\theta(X_i) z_\theta(X_{i+t})^T$$

and w is a so-called 'lag window' chosen so that $w(t) = 1$ for small $|t|$, $w(t) = 0$ for large $|t|$, and w makes a smooth transition from one to the other (see, e. g. Priestly, 1981, pp. 323–324 and 429–435 for a discussion of these issues).

5 Discussion

Some apology should perhaps made for the use of epiconvergence and hypoconvergence, tools that are not part of the working knowledge of most statisticians. One reason is that hypoconvergence clarifies the role played by compactification of the parameter space in Wald-type theorems. Theorem 1 can be stated without reference to a compactification or to other technical means of controlling the oscillations at infinity. Another reason is that there are important consequences that follow from hypoconvergence alone, e. g., Corollary 3. Moreover, these consequences are well-known in optimization theory; once hypoconvergence is established, a wealth of immediate corollaries present themselves.

Theorem 1 differs from Wald-type theorem in another important respect: there is no integrability condition (such as Theorem 2 and Wald (1949) require). This arises from the simple difference in the problems that the randomness is in the denominator in the problem of convergence of Monte Carlo likelihood calculations and in the numerator for problem of consistency of maximum likelihood under repeated sampling, so that for Monte Carlo we need to control an infimum rather than a supremum. Though the difference is trivial it has surprising consequences (surprising to me, at least). The analogy between the Monte Carlo and the repeated sampling problems is very strong, but this one trivial difference makes a huge difference in the regularity conditions that must be imposed to get the result. The question of convergence of Monte Carlo maximum likelihood calculations is essentially resolved. It 'always' works. The only regularity conditions are the minimal amount of continuity required for the topology of the parameter space to have some connection with the probabilities induced by the model. Consistency of maximum likelihood, on the other hand, is plagued by pathological counterexamples like that of Bahadur (1958), and a large literature has been produced about various ways to weaken Wald's integrability condition.

Monte Carlo calculations run into the same difficulty when any randomness appears 'in the numerator' as with the missing data case. Then something like a Wald-type integrability condition must be imposed. Another kind of Monte Carlo calculation where the same need arises is the 'mixture of complete data likelihoods' estimator of the posterior density (Gelfand and Smith, 1990). With such an integrability condition the Monte Carlo approximation to the posterior density will converge to the exact posterior uniformly on compact sets, which implies convergence in total variation of the associated probability distributions.

As mentioned in Section 1 there is no need that the distribution P_ψ from which we sample actually be a distribution in the model. There are good reasons for not using a distribution in the model, what Professor Green called sampling from the 'wrong' model in his discussion of Geyer and Thompson (1992). Other schemes for sampling from the 'wrong' model are given by Sheehan and Thomas (1991) and Geyer (1991). Choosing P_ψ well can make a tremendous difference in the efficiency of sampling (as measured by the variance calculated as described in Section 4), so the choice is important.

'Normalized families of densities' are an important class of statistical models. We now have two interesting properties that hold for the whole class. The Hastings algorithm can be used to simulate realizations from any distribution in the model, and Monte Carlo likelihood approximation can be used to do likelihood-based statistical inference. Since the class is extremely flexible, it allows a very wide scope for modeling and supports the notion of a 'model liberation movement' called for by Professor A. F. M. Smith in his discussion of Geyer and Thompson (1992). There is no need for reasons of mathematical tractability to interfere with using models that are scientifically correct.

Acknowledgements

Conversations with Elizabeth Thompson, Julian Besag, and Michael Newton helped change my focus from exponential families to the general 'normalized families' of Section 1. The whole approach to convergence of optimization problems used in this paper comes from a course taught by Terry Rockafellar in 1990 at the University of Washington using a draft of the book (Rockafellar and Wets, forthcoming). Xiaotong Shen found a mistake in my first proof of Theorem 1. The connection between Monte Carlo likelihood calculations and the posterior density estimator of Gelfand and Smith (1990) was explained to me by Augustine Kong.

References

Attouch, H. (1984) *Variational Convergence of Functions and Operators.* Boston: Pitman.

Bahadur, R. R. (1958) Examples of inconsistency of maximum likelihood estimates. *Sankhyā*, **20**, 207–210.

—— (1971) *Some Limit Theorems in Statistics.* Philadelphia: SIAM.

Chan, K. S. (1991) Asymptotic behavior of the Gibbs sampler. Technical Report No. 294, Department of Statistics, University of Chicago.

Chung, K. L. (1967) *Markov Chains with Stationary Transition Probabilities*, second edition. Berlin: Springer-Verlag.

Gelfand, A. E. and Carlin, B. P. (1991) Maximum likelihood estimation for constrained or missing data models. Research Report 91–002, Division of Biostatistics, University of Minnesota.

Gelfand, A. E. and Smith A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Geyer, C. J. (1990) *Likelihood and Exponential Families.* Ph. D. Dissertation, University of Washington.

Geyer, C. J. (1991) Reweighting Monte Carlo mixtures. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B., to be published.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, **27**, 887–906.

Liu, J., Wong, W. H., and Kong, A. (1991) Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report No. 304, Department of Statistics, University of Chicago.

Nummelin, E. (1984) *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge University Press.

Priestly, M. B. (1981) *Spectral Analysis and Time Series.* London: Academic Press.

Rockafellar, R. T. (1970) *Convex Analysis.* Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J. B. (forthcoming) *Variational Analysis.* New York: Springer-Verlag.

Sheehan, N. and Thomas, A. (1991) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics*, to be published.

Shervish, M. J. and Carlin, B. P. (1990) On the convergence rate of successive substitution sampling. Technical Report, No. 492, Department of Statistics, Carnegie-Mellon University.

Tierney, L. (1991) Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.